Easy Perturbation EEG Algorithm for Spectral Importance (easyPEASI): A Simple Method to Identify Important Spectral Features of EEG in Deep Learning Models

David O. Nahmias
University of Maryland &
US Food and Drug Administration
College Park, Maryland, USA
dnahmias@umd.edu

Kimberly L. Kontson
US Food and Drug Administration
Silver Spring, Maryland, USA
kimberly.kontson@fda.hhs.gov

ABSTRACT

Efforts into understanding neurological differences between populations is an active area of research. Deep learning has recently shown promising results using EEG as input to distinguish recordings of subjects based on neurological activity. However, only about one quarter of these studies investigate the underlying neurophysiological implications. This work proposes and validates a method to investigate frequency bands important to EEG-driven deep learning models. Easy perturbation EEG algorithm for spectral importance (easyPEASI) is simpler than previous methods and requires only perturbations to input data. We validate easyPEASI on EEG pathology classification using the Temple University Health EEG Corpus. easyPEASI is further applied to characterize the effects of patients' medications on brain rhythms. We investigate classifications of patients taking one of two anticonvulsant medications, Dilantin (phenytoin) and Keppra (levetiracetam), and subjects taking no medications. We find that for recordings of subjects with clinically-determined normal EEG that these medications effect the Theta and Alpha band most significantly. For recordings with clinically-determined abnormal EEG these medications affected the Delta, Theta, and Alpha bands most significantly. We also find the Beta band to be affected differently by the two medications. Results found here show promise for a method of obtaining explainable artificial intelligence and interpretable models from EEG-driven deep learning through a simpler more accessible method perturbing only input data. Overall, this work provides a fast, easy, and reproducible method to automatically determine salient spectral features of neural activity that have been learned by machine learning models, such as deep learning.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Applied computing** → *Life and medical sciences*.

KEYWORDS

Deep Learning; Explainable Artificial Intelligence; Electroencephalography (EEG)

This work is licensed under a Creative Commons Attribution-NoDerivs International 4.0 License.

KDD '20, August 23–27, 2020, Virtual Event, USA

© 2020 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-7998-4/20/08.

<https://doi.org/10.1145/3394486.3403289>

ACM Reference Format:

David O. Nahmias and Kimberly L. Kontson. 2020. Easy Perturbation EEG Algorithm for Spectral Importance (easyPEASI): A Simple Method to Identify Important Spectral Features of EEG in Deep Learning Models. In *26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '20)*, August 23–27, 2020, Virtual Event, USA. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3394486.3403289>

1 INTRODUCTION

Deep learning has recently propelled the field of machine learning to new heights. More traditional machine learning models required domain experts to engineer quantitative measurements, or features, which represented the data in meaningful ways. However, deep learning models have introduced automated feature extraction as part of the end-to-end algorithm. By combining the automation of feature extraction and classification, more tailored features of the data that separate groups defined in a classification problem can be found. In applications whose primary objective it is to obtain the highest level of classification, these methods have led to state-of-the-art results without an understanding of what patterns were learned from the data [16]. However, in clinical applications, it may be vitally important to understand the underlying mechanisms of what was automatically learned by the deep learning models [2]. Take, for example, a case in which physiological signals acquired from a patient population and a healthy population are applied to a deep learning model and used for disease detection or diagnosis. By understanding the underlying physiological differences between populations classified by deep learning models, it is possible to discover potentially new biomarkers of disease or elucidate the progression and impact of a disease on those physiological functions.

Electroencephalography (EEG) is one modality that captures such physiological signals and is used both in research and clinical settings to detect or diagnose medical conditions [17]. EEG measures brain electrical signals from the scalp in a typically non-invasive and relatively inexpensive way and has been shown to be a promising neuro-monitoring technique to identify a wide range of neuro-physiological biomarkers for many different diseases [6]. However, EEG presents complex data that has a low signal to noise ratio (SNR), thus making the data difficult to interpret. More difficult yet is understanding how to best quantify EEG using complex signal processing and mathematical methods. Because of these challenges, deep learning has recently shown promising results using EEG directly as input to distinguishing individuals based on neurological activity. However, only about one quarter of these

Figure 1: easyPEASI flowchart. Blue semi-rounded items represent pre-computed elements, purple rectangular items are intermediate elements, orange rounded items are intermediate results, and the green circle represents the last step and result. Elements in dotted area are repeated for each $1, 2, \dots, k$ samplings.

studies investigate the underlying physiological implications on which these models are based [14].

Much like the early development of deep learning models, methods to understand and visualize these models have mainly developed and focused in the image processing domain [22]. With the adoption of these models to EEG, some of these investigation methods have also been adapted. The most common method of model inspection analyzes model weights learned. These analyses are usually limited to the first layer of the model because of the limited interpretability of subsequent filter layers [20]. Another common approach has been to investigate the model activations on representative inputs [15]. Additionally, the ideas of input perturbation and input occlusion have been explored in limited context [14]. Using established EEG-based deep learning models, an input-perturbation network-perturbation correlation map technique was developed. These methods combine both input perturbation and model weight analysis to determine the correlation between input and network characteristics [16]. When applying occlusion-based methods in image processing, parts of the input image are occluded and the effect on the network output is analyzed to determine the most important area of the image to the model. Some research in analyzing EEG-based deep learning models have generated images that represent the EEG and subsequently applied occlusion methods on those images [19]. Although previous research is promising and allows for better understanding learned features in deep learning models, most of these methods require in depth analysis of model parameters or activations and can be difficult to interpret. Given the widespread application of deep learning in many different fields and the potential for interpretation by individuals who do not have expert knowledge in these models, the development of a simple, easily interpretable, automated method that manipulates EEG directly to determine important learned EEG features from deep learning models is needed.

This work seeks to address this need and expand upon previous methods by applying perturbation methods in image processing applications to EEG directly. There is a long history of analyzing frequency bands in EEG to understand underlying neurological mechanisms and most work in EEG has focused on understanding frequency bands of importance [11]. We introduce a novel method of applying frequency domain perturbations to EEG to understand how important specified frequency bands are on the identifying populations based solely on EEG. The ability to decipher important features in the EEG has many potential applications. Specifically,

deep learning can be used to identify subjects' medication status, specifically anticonvulsant medications. To develop more informative EEG based pharmacodynamic models, it is important to determine which, if any, specific frequency bands are the underlying bases for the differences in anticonvulsant effects.

The aims of this work are to (1) introduce a novel, simpler input perturbation method for deep-learning classification problems that allows for the determination of important frequency band features in a clinical application, (2) validate this novel approach by comparing those features identified as important between our novel approach and other established approaches, and (3) apply our proposed method to patient medication status classification problems to determine which EEG frequency bands are important in distinguishing individuals taking different anticonvulsant medications or individuals taking no medications.

2 METHODS

We first introduce the novel algorithm and applied statistical methods to determine important frequency bands for a given classification problem. We next describe the specific deep learning model used to validate the novel method and the EEG data set used for validation and application of the novel method.

2.1 easyPEASI: Easy Perturbation EEG Algorithm for Spectral Importance

At a high level, the proposed easy perturbation EEG algorithm for spectral importance (easyPEASI) is shown visually in Figure 1. easyPEASI requires that k -fold samplings of the EEG data and associated labels be obtained. These k sub-sets are used to train k binary classification deep learning models. With input data in the form of a collection of EEG channels recordings across time, the aim of this method is to identify important frequency bands that distinguish the labeled data. This is accomplished by introducing perturbations (i.e. noise) to the recordings to specific frequency bands for all recorded channels for each subject. Entire frequency bands were analyzed rather than each individual frequency since these bands have known clinical meanings and power within each band may not be equally and consistently distributed across all frequencies for all recordings. Investigating single frequencies may also smear results when analyzing results across many recordings and obscure meaningful insights.

Using the Discrete Fourier Transform (DFT) for a real-valued signal, rFFT (real Fast Fourier Transform) of $EEG[channel]$,

$$\Omega = \text{rFFT}(EEG[channel]) = \begin{cases} [y[0], \text{Re}(y[1]), \text{Im}(y[1]), \dots, \text{Re}(y[n/2])], & n \text{ even} \\ [y[0], \text{Re}(y[1]), \text{Im}(y[1]), \dots, \text{Re}(y[n/2]), \text{Im}(y[n/2])], & n \text{ odd} \end{cases} \quad (1)$$

where

$$y[j] = \sum_{k=0}^{n-1} EEG[channel, k] \exp\left(-i \cdot \frac{j \cdot k \cdot 2\pi}{n}\right) \quad (2)$$

for $j = \{0, \dots, n-1\}$, we obtain the frequency domain representation of $EEG[channel]$ [21].

The central manipulation in easyPEASI of the EEG data is the replacement of the real and complex DFT coefficients of the targeted frequency band in line 6 of the *PerturbEEGBands* function. These coefficients are replaced with random numbers generated from a Normal distribution, $\mathcal{N}(\mu, \sigma)$, where μ and σ are the mean and standard deviation (SD), respectively, of the real and complex DFT coefficients of Equation (1) in the targeted frequency band. Once replaced, the inverse DFT, irFFT (inverse rFFT), is computed to obtain the final perturbed EEG channel. This specific spectral manipulation affects the spectral power dynamics within the frequency band, the phase coupling of the frequencies within the band and with other frequencies, and the non-stationarity of the frequencies in the band present in signal.

From this we describe the steps for our proposed method. Two parameters of choice that further need to be defined are frequency bands of interest and the number of k -fold data samplings. The frequency bands commonly analyzed in the EEG are as follows: δ (Delta) : 1–4Hz, θ (Theta) : 4–8Hz, α (Alpha) : 8–12Hz, μ (Mu) : 12–16Hz, β (Beta) : 16–25Hz, γ (Gamma) : 25–40Hz [17]. These frequency band bounds were used as the f_Low and f_High for each subject's EEG data as the inputs to the perturbation function, defined in Line 3 in *easyPEASI*. Further, we used a $k = 10$ -fold cross-validation factor implementing this algorithm. In our outline of *easyPEASI* below, boldface is used for keywords, italics represent functions, underlined italics functions that are defined as part of *easyPEASI*, and all others are variables.

Using random subsets of the data, k deep learning models were trained and training classification accuracies for those models were calculated. Perturbations were then applied in a specific frequency band (*PerturbEEGBands* function) to those same random subsets of data and classification accuracies were calculated using the k previously defined deep learning models. The k classification accuracies from the training (unperturbed) data and k classification accuracies from the perturbed data were used as input in the Kruskal-Wallis test. The result of the Kruskal-Wallis test determined the statistical significance of the difference between the sets of accuracies, P in line 8 of the *BandImportance* function [5]. The mean of the k training accuracies and the mean of the k perturbed accuracies are also compared by measuring their difference, Δ in line 9 of the *BandImportance* function.

Though $k = 10$ was used here, other values $k > 5$ can be used to meet Kruskal-Wallis test requirements. The null hypothesis of the Kruskal-Wallis test is that the medians of group of samples are

the same implying that the distributions of the measurements are consistent. The H statistic was computed on ranks and a P -value obtained using a Chi-Square distribution by $Pr[\chi^2(df-1) \geq H]$, where $df = k$ is the degrees of freedom. Therefore, if the perturbed data classification results were not similar to the original training classification results and significant P -values were found, then the effect of the perturbation on the frequency band was informative and important in the classification task. We used thresholds of $P < .01$ and $P < .001$ as varying levels of significance, though other P -value thresholds could be used.

Algorithm easyPEASI: Easy perturbation EEG algorithm for spectral importance

```

1: function PERTURBEEGBANDS(EEG, F_Low, F_High)
2:   for channel  $\in$  EEG do
3:      $\Omega \leftarrow \text{rFFT}(EEG[channel])$ 
4:      $\mu \leftarrow \text{Mean}(\Omega[F\_Low, F\_High])$ 
5:      $\sigma \leftarrow \text{SD}(\Omega[F\_Low, F\_High])$ 
6:      $\Omega[F\_Low, F\_High] \leftarrow \mathcal{N}(\mu, \sigma)$ 
7:      $EEG[channel] \leftarrow \text{irFFT}(\Omega)$ 
8:   return EEG
9:
10: function BANDIMPORTANCE(k, F_Low, F_High)
11:   for i  $\leftarrow$  0 to k - 1 do
12:     EEG, Labels  $\leftarrow$  LoadTrainData(i)
13:     Model  $\leftarrow$  LoadTrainedDeepModel(i)
14:     PerturbedEEG  $\leftarrow$  PerturbEEGBands(EEG, F_Low, F_High)
15:     TrainAcc[i]  $\leftarrow$  EvalModel(Model, EEG, Labels)
16:     PerturbAcc[i]  $\leftarrow$  EvalModel(Model, PerturbedEEG, Labels)
17:
18:    $P \leftarrow \text{KruskalWallisTest}(\text{TrainAcc}, \text{PerturbAcc})$ 
19:    $\Delta \leftarrow \|\text{Mean}(\text{TrainAcc}) - \text{Mean}(\text{PerturbAcc})\|$ 
20:   return P,  $\Delta$ 
21:
22: procedure EASYPEASI()
23:   k  $\leftarrow$  10 ▷ k-fold samplings
24:   BandFreqLim  $\leftarrow$  [1,4,8,16,25,40] ▷ Hz
25:   for b  $\leftarrow$  0 to length(BandFreqLim)-2 do
26:     F_Low  $\leftarrow$  BandFreqLim[b]
27:     F_High  $\leftarrow$  BandFreqLim[b+1]
28:     P,  $\Delta \leftarrow$  BandImportance(k, F_Low, F_High) ▷ Results

```

2.2 Trained deep convolutional neural network

As discussed, easyPEASI requires that trained EEG-based deep learning model and corresponding labeled training EEG data be available. We used a variant of a deep convolutional neural network (DCNN) model introduced in previous research and further modified for data used in this study (Appendix A) [16]. For each binary classification, population groups were evenly balanced with recordings from each of the two labels. This was done so that there would be equal prevalence of each condition. 90% of the available data were randomly sampled and used to train the DCNN (line 2 in Train Deep Models). This process was repeated k times to obtain k trained models using different random sub-sets of data. The application was run on an NVIDIA GTX-1080Ti for training through Python 3.5 and Torch [16], [10].

Table 1: Population demographics of data set across recordings with EEG pathology status, and medication profiles of interest.

	All recordings	Normal EEG	Abnormal EEG	Taking Dilantin	Taking Keppra	No medications
Total n	35,370	6,001	15,347	718	898	5,181
Normal EEG	6,001	6,001	0	179	175	1,203
Abnormal EEG	15,347	0	15,347	320	264	1,285
EEG status not obtained	14,022	0	0	291	459	2,693
Male	15,106	2,371	7,088	344	378	2,017
Female	16,434	3,222	7,289	320	465	1,895
Sex not obtained	3,830	408	970	54	55	1,269
Age (Mean $\pm$ SD years)	50.23 $\pm$ 19.15	41.14 $\pm$ 18.38	52.90 $\pm$ 19.04	48.36 $\pm$ 16.72	48.04 $\pm$ 18.98	44.55 $\pm$ 20.11
(# recordings w/ age)	($n = 33,021$)	(5,865)	($n = 15,027$)	($n = 692$)	($n = 862$)	($n = 4,142$)
Unique subjects	13,486	4,058	7,432	456	507	2,530

Algorithm Train Deep Models

```

1:  $k \leftarrow 10$                                  $\triangleright$  k-fold samplings
2: Ratio  $\leftarrow 0.9$                          $\triangleright$  Training set proportion
3: Data, Labels  $\leftarrow LoadData()$ 
4: for  $i \leftarrow 0$  to  $k - 1$  do
5:   TrainData, TrainLabels  $\leftarrow SampleData(Data, Labels, Ratio)$ 
6:   Model  $\leftarrow TrainDeepModel(TrainData, TrainLabels)$ 
7:   SaveData(TrainData, TrainLabels)
8:   SaveModel(Model)

```

2.3 EEG data

EEG data from the Neural Engineering Data Consortium (NEDC) EEG Corpus from Temple University Hospital (TUH) version 1.0.0 were used as input to the DCNN. This is an unprecedented real-world data set composed of thousands of subject EEG recordings collected under conditions less constrained than those typically employed in laboratory research settings [3]. This data set contains 35,370 EEG recordings sessions (1.1TB) from 13,486 subjects. Each session was accompanied by a report containing clinical impressions and patient characteristics, including age and sex, as well as the medications being taken by the patient at the time of recording. It primarily consists of data from recordings of subjects with EEG abnormalities, with many subjects taking anticonvulsant medications. Thus, the TUH data set provides the means to explore EEG pathologies and medication state classifications by evaluating the neurophysiological signals of thousands of subjects. To investigate the different frequency band importance of EEG-based classifications in these populations we applied several processing steps to the EEG data and corresponding clinical reports from the TUH data set.

Using this data set, the two most common anticonvulsant medications taken alone based on the clinical reports were identified: Dilantin (generic phenytoin) and Keppra (generic levetiracetam) (Table 1). Any subject with more than one medication listed in the clinical notes was excluded from the evaluation to reduce results related to drug interaction effects. Furthermore, given the clinical nature of the data set and to avoid confounds from clinically-determined normal and abnormal EEGs, these classifications were

performed separately on recordings deemed normal EEG and abnormal EEG. A subset of patient clinical reports within the TUH data set had been previously reviewed and labeled by curators of this database to be clinically normal or abnormal [7]. The definitions of normal and abnormal used by the TUH generally followed methods neurologists use to identify abnormalities [8]. These labels were applied broadly to recordings where the clinical reports identified an abnormality in the EEG and do not distinguish between different abnormalities. An automated labeling method previously developed and validated based on textual analysis of the clinical reports was applied to categorize all available recordings as normal or abnormal [9]. Applying this method to the entire TUH data set resulted in a total of 6,001 EEG recordings with normal EEG across 4,058 unique subjects and 15,347 EEG recordings with abnormal EEG across 7,432 unique subjects.

Because the recordings in the TUH data set were acquired over several years and with different EEG systems, a subset of channels following a standard 10-20 electrode montage were used: Fp1, Fp2, F3, F4, C3, C4, P3, P4, O1, O2, F7, F8, T3, T4, T5, T6, Fz, Cz, Pz. Our analyses were applied to five minutes of EEG recordings across those 19 channels identified above [7]. If another sampling rate was used, recordings were down-sampled to 100Hz. The data were then filtered through a 0.5Hz to 50Hz band pass filter, and re-referenced using a common average. All pre-processing and analyses were done in Python using a combination of available packages, published code from previous works, and custom developed code [12], [16] (available at github.com/dbp-osel).

3 RESULTS**3.1 Validation: Spectral importance of EEG pathology detection**

We validated easyPEASI though classifying EEG pathologies, normal EEG recordings against abnormal EEG recordings, since previous works investigating deep learning models that used the TUH data classified EEG pathology states [16]. This allows for us to compare results from the same classification using a similar data set. We report the mean and standard deviation of the 10-fold cross validation training accuracy results as the baseline for how well the deep learning model learned classification results as a line and

Figure 2: easyPEASI results for models detecting EEG pathologies. Each bar represents the classification accuracy from the DCNN model when a particular frequency band was perturbed using the easyPEASI. Line and shaded area at $80.99\% \pm 2.22$ are mean and standard deviation of training accuracies with no perturbed data. Kruskal-Wallis P -value significance thresholds set at: n.s.: not significant, *: $P < .01$, **: $P < .001$.

Table 2: Classification accuracies (Mean $\pm$ Standard deviation %) for subject EEG pathologies.

Classification	n -train (n -test)	Train accuracy (%) (Test accuracy %)
Normal EEG vs. Abnormal EEG	10,802 (1,200)	80.99 ± 2.22 (80.40 ± 2.41)

shaded region in Figure 2. With 6,001 available normal EEG recordings and 15,347 abnormal EEG recordings, balancing these two populations yields an n -total = 12,002 and using 90% of the data for training gives n -train=10,802. These training models generalized well through 10-fold cross-valued testing results with n -test=1,200. The training and testing results of this classification are shown in Table 2. We further report the differences, Δ , and P -values from Kruskal-Wallis tests between the results obtained from the training data and perturbed data where we denote *: $P < .01$, **: $P < .001$, or n.s.: not significant in Figure 2.

As in previous work identifying band importance, we find that the Delta and Theta bands were most distinctive between populations when classifying EEG pathologies [16], instilling confidence that our novel method to identifying these bands is valid.

3.2 Application: Anticonvulsants effects on brain rhythms

Given the validation of this novel algorithm, we next report the results of applying easyPEASI to determining those frequency bands most important to medication state classifications for anticonvulsants. These training data sets and trained models are made available for rapid reproducibility (Appendix A). In this work we focus on training accuracies for comparison to perturbed data accuracy values found with easyPEASI. We began by comparing recordings

Table 3: Classification accuracies (Mean $\pm$ Standard deviation %) for recordings of subjects taking one of two anticonvulsants, either Dilantin or Keppra, with normal or abnormal EEGs.

Classification	n -train (n -test)	Train accuracy (%) (Test accuracy %)
Dilantin vs. Keppra with Normal EEG	316 (34)	80.19 ± 13.86 (59.12 ± 7.49)
Dilantin vs. Keppra with Abnormal EEG	476 (52)	83.91 ± 3.48 (60.39 ± 5.03)

of subjects taking either only Dilantin or only Keppra. We further separated subjects based on whether they had abnormalities present in their EEG recordings. Using 90% of the data for training, recordings of subjects with normal EEG taking Dilantin vs. Keppra we report training and testing accuracies in Table 3 and show easyPEASI results in Figure 3. These models obtained results above 50%, which is considered chance for a binary classification. Applying easyPEASI to the training data, we show the importance of each band in the two classifications identifying anticonvulsants.

In the case of normal EEG recordings, no frequency bands were found to be significant in differentiating between the effects of the two anticonvulsants Dilantin and Keppra, though we note the Alpha band was found to have a $\Delta = 16.80\%$, which was 10% larger than any other frequency band. Further, for abnormal EEG recordings we found Theta, Alpha, Delta and Beta bands were most important for this classification.

We then evaluated the bands' importance to distinguishing recordings of subjects taking either one of the two anticonvulsants against subjects taking no medications. Again, classifications and easyPEASI were applied separately for abnormal and normal EEG recordings. Table 4 shows the mean and standard deviation of the training

Figure 3: easyPEASI results for models comparing the differences between anticonvulsants normal and abnormal EEG recordings. Each bar represents the classification accuracy from the DCNN model when a particular frequency band was perturbed using the easyPEASI. Lines and shaded areas at (a) $80.19\% \pm 13.86$ and (b) $83.91\% \pm 3.48$ are means and standard deviations of training accuracies with no perturbed data. Kruskal-Wallis P -value significance thresholds set at: n.s.: not significant, *: $P < .01$, **: $P < .001$.

Table 4: Classification accuracies (Mean $\pm$ Standard deviation %) for recordings of subjects taking one of two anticonvulsants and subjects taking no medications.

Classification	n -train (n -test)	Train accuracy (%) (Test accuracy (%))
Dilantin vs. No medications with Normal EEG	324 (34)	82.22 ± 8.24 (64.12 ± 8.19)
Keppra vs. No medications with Normal EEG	316 (34)	85.82 ± 5.83 (64.71 ± 6.31)
Dilantin vs. No medications with Abnormal EEG	576 (64)	84.36 ± 7.21 (68.75 ± 6.29)
Keppra vs. No medications with Abnormal EEG	476 (52)	82.69 ± 5.86 (68.85 ± 6.87)

and testing accuracies obtained, with easyPEASI results in Figure 4 and Figure 5. Table 4 shows that testing classification results were above chance. Applying easyPEASI, the important bands for these classifications are shown in Figure 4 for normal EEG recordings and Figure 5 for abnormal EEG recordings.

The results for recordings of subjects taking a single medication and subjects taking no medications show that for normal EEG recordings, both anticonvulsants affected the Alpha band most, followed by the Theta band. The Beta band was also found to be significant for recordings of subjects with normal EEG taking Keppra against subjects taking no medications. In the cases of abnormal EEG recordings, both classifications comparing subjects taking anticonvulsants and subjects taking no medications, the Delta band was found to be most significantly affected. However, the Theta and Alpha bands were only significant when distinguishing recordings of subjects with abnormal EEG taking Dilantin against those taking no medications.

4 DISCUSSION

In the current study, we introduce and validate a method (easyPEASI) to investigate deep learning models that use EEG as their input to identify important frequency bands in a given binary classification problem and provide an example of how this algorithm can be applied to real clinical classifications problems. Unlike previous methods that have applied some type of manipulation on the input data and associated those manipulations to changes in deep learning network weights and activations, our method only requires applications of perturbations to the input data to determine

Figure 4: easyPEASI results for models comparing the differences for recordings of subjects with normal EEG taking one anticonvulsant and subjects with normal EEG taking no medications. Each bar represents the classification accuracy from the DCNN model when a particular frequency band was perturbed using the easyPEASI. Lines and shaded areas at (a) $82.22\% \pm 8.24$ and (b) $85.82\% \pm 5.83$ are means and standard deviations of training accuracies with no perturbed data. Kruskal-Wallis P -value significance thresholds set at: n.s.: not significant, *: $P < .01$, **: $P < .001$.

frequency band importance. Our work focuses on frequency bands because of its prominence in prior work to investigate underlying neuro-physiology through EEG [11].

We first validated our method by identifying frequency bands most important in classifying EEG pathologies, as done with other methods [16]. With these results we see that using easyPEASI to investigate EEG-driven deep learning models shows promise in identifying important frequency bands in classification tasks. This method allows for the investigation of important spectral characteristics of EEG learned by deep learning models without the need to directly investigate the deep learning model itself. The method introduced and validated in this work is simpler than previous methods since it does not require investigating the deep learning model and rather only relies on manipulation of the input data. The interpretation of the results is also based upon a commonly used, well-known statistical method, the Kruskal-Wallis test. Therefore, easyPEASI may be more accessible to researchers in the field of neuroscience that may not be experts in deep learning but often require explainable artificial intelligence and explainable models to understand clinical implications. The spectral perturbation that the *PerturbEEGBand* function in easyPEASI applies to the EEG signal is simple in nature and affects particular properties of the signal. Replacing the real and complex coefficients of the DFT with normally

distributed random values affects the peaks and troughs of the spectral power profile. Further, when applying the perturbation to the data, the phase coupling and non-stationarity of the EEG signal is affected. Therefore, it is possible that any combination of these properties of each frequency band may have been most influential when significant bands were identified by easyPEASI. Along with varying the number of data samplings k used or frequency bands investigated, different perturbations that manipulate other specific properties of frequencies may be investigated in future works with easyPEASI.

Identification of EEG pathologies through deep learning models using solely EEG as inputs achieved accuracies above 80% using about 3,000 EEG recordings [16]. Our automated EEG pathology labeling method allowed for classification of EEG pathologies with deep learning using 12,002 EEG recordings, the largest EEG data training and testing sets to classify EEG pathologies. Using this data set also achieved testing accuracy greater than 80% and with easyPEASI identified the same bands important to the classification (Figure 2). These levels of testing accuracy allow for deep learning models for classification to be potentially used in diagnostic applications [18]. However, these levels of accuracy have yet to be achieved using deep learning models to identify subject's medications using EEG, but have shown the ability to distinguish these populations

Figure 5: easyPEASI results for models comparing the differences for recordings of subjects with abnormal EEG taking one anticonvulsant and subjects with abnormal EEG taking no medications. Each bar represents the classification accuracy from the DCNN model when a particular frequency band was perturbed using the easyPEASI. Lines and shaded areas at (a) $84.36\% \pm 7.21$ and (b) $82.69\% \pm 5.86$ are means and standard deviations of training accuracies with no perturbed data. Kruskal-Wallis P -value significance thresholds set at: n.s.: not significant, *: $P < .01$, **: $P < .001$.

above chance (50%). This means that these models generalized to unseen data to a certain degree and identified meaningful patterns in the data. Concurrent research outside the scope of this work is in progress that analyses and discusses the significance and impact of these classification results (Table 3 and Table 4).

As a novel application of easyPEASI, we investigated EEG recordings from individuals taking different anticonvulsant medications and no medications. To our knowledge, this kind of analysis into neurophysiological differences for this drug class is not well researched. Here, we discuss the results of the easyPEASI method and correlate these results with expectations based on clinical literature. For normal EEG recordings, though not statistically significant, the largest differences in effects between medications were found to be in the Alpha band (Figure 3a). Comparing recordings of subjects with normal EEG taking a single medication (i.e. Keppra or Dilantin) to subjects taking no medications, the Alpha band was found to be significantly different (Figure 4). With the Alpha band the most important band comparing normal EEG recordings, it seems that anticonvulsants affect this band in a similar manner when compared to recordings of subjects taking no medications since it is not significant when compared to each other (Figure 3a). Similarly, the Theta band seems to be equally important (Figure 4) and comparably affected by these medications (Figure 3a). Finally,

we note that though the Beta band was found to be important in only classifying recordings of subjects with normal EEG taking Keppra against subjects taking no medications (Figure 4b), the Beta band was not found to be important classifying the two medications for normal EEG recordings (Figure 3a).

For abnormal EEG recordings, results showed both anticonvulsant medications affected the Delta band (Figure 3a and Figure 5). Further, the Theta and Alpha bands were found to be important when classifying recordings of subjects with abnormal EEGs taking Dilantin against those taking no medications (Figure 5a). These results are consistent with known effects of EEG pathologies [16]. Delta and Theta bands have been important for identifying EEG pathologies, so it is expected that the medication aimed to treat epilepsy, one of the most common pathologies, would affect these bands for abnormal EEG recordings [8]. When comparing these medications to each other, Dilantin's mechanism of action, like most anticonvulsant medications, is primarily attributed to modulating voltage-gated sodium and calcium channels and enhancement of GABA_A (gamma aminobutyric acid, type A) [13]. However, Keppra's mechanism of action is less understood but is thought to be through binding to the synaptic vesicle protein SV2A [1]. With these differences, our finding that the Beta band was uniquely differently affected by these two medications (Figure 3b) is consistent

with known neurophysiology since the Beta band is known to be affected by GABA and is only known to be directly affected by one of these medications [4]. Further, since they have different mechanisms of action, though they may both be affecting the Delta bands, they may be doing so differently since we observe the Delta band was important when comparing anticonvulsants to each other (Figure 3b). However, we also found the Theta and Alpha bands to be important when distinguishing these medications from each other for abnormal EEG recordings (Figure 3b) and were only found to be important comparing recordings of subjects taking Dilantin to those taking no medications (Figure 5a).

Overall, this work provides a fast, easy, and reproducible method to automatically determine salient spectral features of neural activity that have been learned by machine learning models, such as deep learning. A limitation of this method, however, is that the perturbations were applied to all channels together. Future development of extensions to easyPEASI could be done to identify brain regions that are important in classification tasks. Regarding the applications investigated in this work, future research may investigate the difference between subjects with abnormal EEG taking anticonvulsants against subjects with normal EEG taking no medications to potentially determine anticonvulsants' efficacy.

ACKNOWLEDGMENTS

This work was funded by Center for Devices and Radiological Health, Office of Science and Engineering Laboratories, Division of Biomedical Physics.

DISCLAIMER

The mention of commercial products, their sources, or their use in connection with material reported herein is not to be construed as either an actual or implied endorsement of such products by the Department of Health and Human Services.

REFERENCES

- [1] Bassel Abou-Khalil. 2008. Levetiracetam in the treatment of epilepsy. *Neuropsychiatric disease and treatment* 4, 3 (2008), 507–523. <https://doi.org/10.2147/ndt.s2937>
- [2] Jean-Marc Fellous, Guillermo Sapiro, Andrew Rossi, Helen Mayberg, and Michele Ferrante. 2019. Explainable Artificial Intelligence for Neuroscience: Behavioral Neurostimulation. *Frontiers in Neuroscience* 13 (2019), 1346. <https://doi.org/10.3389/fnins.2019.01346>
- [3] A. Harati, S. Lopez, I. Obeid, M. Jacobson, S. Tobochnik, and J. Picone. 2014. The TUH EEG Corpus: A Big Data Resource for Automated EEG Interpretation. In *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium* (2014), 1–5.
- [4] O. Jensen, P. Goel, N. Kopell, M. Pohja, R. Hari, and B. Ermentrout. 2005. On the human sensorimotor-cortex beta rhythm: Sources and modeling. *NeuroImage* 26, 2 (2005), 347–355. <https://doi.org/10.1016/j.neuroimage.2005.02.008>
- [5] William H. Kruskal and W. Allen Wallis. 1952. Use of Ranks in One-Criterion Variance Analysis. *J. Amer. Statist. Assoc.* 47, 260 (1952), 583–621.
- [6] Chi Qin Lai et al. 2018. Literature survey on applications of electroencephalography (EEG). *AIP Conference Proceedings* 2016, 020070 (2018).
- [7] S. Lopez. 2017. Automated Identification of Abnormal EEGs. *Temple University* (2017).
- [8] S. López, G. Suarez, D. Jungreis, I. Obeid, and J. Picone. 2015. Automated Identification of Abnormal Adult EEGs. In *IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*. <https://doi.org/10.1109/SPMB.2015.7405423>
- [9] David O. Nahmias, Kimberly L. Kontson, David A. Soltysik, and F. Civillico, Eugene. 2019. Consistency of Quantitative Electroencephalography Features in a Large Clinical Data Set. *Journal of Neural Engineering* 16, 066044 (2019). <https://doi.org/10.1088/1741-2552/ab4af3>
- [10] F. Nasse, C. Thureau, and G. A. Fink. 2009. A review of classification algorithms for EEG-based brain-computer interfaces: a 10 year update. *International Conference on Computer Analysis of Images and Patterns* (2009), 83–90.
- [11] Jennifer J. Newson and Tara C. Thiagarajan. 2019. EEG Frequency Bands in Psychiatric Disorders: A Review of Resting State Studies. *Frontiers in Human Neuroscience* 12 (2019), 521. <https://doi.org/10.3389/fnhum.2018.00521>
- [12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
- [13] M. Rogawski and W. Löscher. 2004. The neurobiology of antiepileptic drugs. *Nature Review Neuroscience* 5 (2004), 553–564. <https://doi.org/10.1038/nrn1430>
- [14] Yannick Roy et al. 2019. Deep learning-based electroencephalography analysis: a systematic review. *Journal of Neural Engineering* 16, 051001 (2019). <https://doi.org/10.1088/1741-2552/ab260c/>
- [15] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K. Müller. 2017. Evaluating the Visualization of What a Deep Neural Network Has Learned. *IEEE Transactions on Neural Networks and Learning Systems* 28, 11 (Nov 2017), 2660–2673. <https://doi.org/10.1109/TNNLS.2016.2599820>
- [16] R. Schirrmester, L. Gemein, Eggensperger, F. Hutter, and T. Ball. 2017. Deep learning with convolutional neural networks for decoding and visualization of EEG pathology. *The IEEE Signal Processing in Medicine and Biology Symposium* (December 2017).
- [17] Donald L. Schomer and Fernando H. Lopes da Silva. 2010. *Niedermeyer's Electroencephalography: Basic Principles, Clinical Applications, and Related Fields* (6 ed.). Lippincott Williams & Wilkins.
- [18] J. Shen, C. Zhang, B. Jiang, J. Chen, Z. Song, J. and Liu, and W. K. Ming. 2019. Artificial Intelligence Versus Clinicians in Disease Diagnosis: Systematic Review. *JMIR medical informatics* 7(3), e10010 (2019). <https://doi.org/10.2196/10010>
- [19] Pierre Thodoroff, Joelle Pineau, and Andrew Lim. 2016. Learning Robust Features using Deep Learning for Automatic Seizure Detection. *CoRR abs/1608.00220* (2016). <http://arxiv.org/abs/1608.00220>
- [20] Michel J. A. M. van Putten, Sebastian Olbrich, and Martijn Arns. 2018. Predicting sex from brain rhythms with deep learning. *Nature: Scientific Reports* 8, 3069 (2018).
- [21] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2019. SciPy 1.0—Fundamental Algorithms for Scientific Computing in Python. *arXiv e-prints*, Article arXiv:1907.10121 (Jul 2019), arXiv:1907.10121 pages. [arXiv:cs.MS/1907.10121](https://arxiv.org/abs/1907.10121)
- [22] Matthew D. Zeiler and Rob Fergus. 2014. Visualizing and Understanding Convolutional Networks. In *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 818–833. https://doi.org/10.1007/978-3-319-10590-1_53

A ONLINE RESOURCES FOR REPRODUCIBILITY

Software and scripts for reproducibility can be found at our GitHub Repository: github.com/dbp-osel.

Software implementation for easyPEASI and analysis of this work are at: github.com/dbp-osel/easyPEASI. This repository also includes saved trained deep learning models for all classifications reported here. For the application area of medication classifications, data and labels are also made available. These resources allow for fast reproducibility of results presented in this work.

For other classifications using the TUH EEG Corpus to reproduce results obtained here (e.g. larger EEG pathology detection data set) several pre-processing steps need to be applied the data before these deep learning models can be trained and subsequently applied to easyPEASI. These pre-processing tools and data are available as part of work previously published: github.com/dbp-osel/qEEG-consistency [9].

The open source EEG data is available from Temple University Health at www.isip.piconepress.com/projects/tuh_eeg [3].